The Breslow Estimator of the Nonparametric Baseline Survivor Function in Cox's Regression Model

Some Heuristics

James A. Hanley

Abstract: Most survival analysis textbooks explain how the hazard ratio parameters in Cox's life table regression model are estimated. Fewer explain how the components of the nonparametric baseline survivor function are derived. Those that do often relegate the explanation to an "advanced" section and merely present the components as algebraic or iterative solutions to estimating equations. None comment on the structure of these estimators. This note brings out a heuristic representation that may help to de-mystify the structure.

(*Epidemiology* 2008;19: 101–102)

Most survival analysis textbooks aimed at nonstatisticians describe how the hazard ratio parameters in Cox's life table regression model[1] are estimated, but few explain how the components of the nonparametric baseline survivor function are derived. Those that do (such as the book by Collett[2]) tend to relegate the description to an "advanced" section, where the components are merely presented as algebraic or iterative or approximate solutions to estimating equations. None comment on the structure of these estimators.

In what follows, I give a heuristic explanation of how these estimators work. By publicizing this explanation, I hope that it will eventually make its way into textbooks. I also advance a possible explanation for an "even more surprising" observation made by Breslow in his contribution[3] to the discussion of Cox's 1972 paper.

The structure is most readily seen with the Product-Limit step-function version of $\hat{S}_0(t)$. Denote the fitted vector of regression coefficients by $\hat{\beta}$. Consider risk set i, with d_i failures at time t_i. Suppose the n_i risk set members have covariate vectors $z_{i,1}$, $z_{i,2}, \ldots, z_{i,ni}$, such that $\sum_j \exp[\hat{\beta} z_{ij}] = n'_i$, say. The choice of n'_i to denote the sum of hazard ratios will become obvious below.

In his contribution to the discussion of Cox's paper, Breslow[3] suggested estimating $S_0[t]$ using "an exact analogue of the Kaplan-Meier estimate," namely the product

$$\hat{S}_0(t) = \prod_{i:t_i < t} (1 - \hat{\pi}_i)$$

where, in the notation used here,

$$\hat{\pi}_i = \frac{d_i}{n'_i}$$

He explained that the "expression for the $\hat{\pi}_i$ can also be obtained as a first-order approximation to the estimate suggested by Cox, and as an approximation to the estimate derived from the distinct discrete time model of Kalbfleisch and Prentice.[4]" Since then, the textbooks and software documentation that have covered estimators of $S_0[t]$ have simply repeated this formula, or the more complex one by Kalbfleisch and Prentice[4] that leads to a closed form expression only when $d_i = 1$. None has remarked on the structure of Breslow's approximation.

The structure is more easily seen if one considers a dataset where there is just one, binary, covariate z, with $z = 1$ if the subject is male and 0 if female. One could estimate $S_0[t]$, the survival function for females, by restricting the classic Kaplan-Meier estimator to the survival times for the females in the sample. But why not use the proportional hazards model to synthetically augment the sample? Why not "transmute the males into females"[5] and estimate $S_0[t]$, by the Kaplan-Meier or Nelson-Aalen method, from a single dataset consisting of both the females and the "female-equivalents" of the males?

Operationally, this translates into the following epidemiologic counting: if $\exp[\hat{\beta} \times 1] = 1.5$, say, then a risk set i that consists of say 50 females, each with a (relative) hazard of $\exp[\hat{\beta} \times 0] = 1$, and 60 males (so, $n_i = 110$ in all) is equivalent to a risk set that consists of

Submitted 16 April 2007; accepted 25 July 2007.

From the Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada.

Supported by grants from The Natural Sciences and Engineering Research Council of Canada and Le Fonds Québecois de la recherche sur la nature et les technologies.

Correspondence: James A. Hanley, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Ave. West, Montreal, Quebec H3A 1A2, Canada. E-mail: james.hanley@mcgill.ca.

ISSN: 1044-3983/08/1901-0101
DOI: 10.1097/EDE.0b013e31815be045

$$n'_i = 50 \times 1 + 60 \times 1.5 = 140 \text{ “females”}$$

Thus the estimate of π_i is $d_i/140$, just as if it were performed in a standard Kaplan-Meier calculation in a “females only” risk set with an “effective size” of 140. The $d_i/140$ is also used in the sum that forms the Nelson-Aalen estimator of the integrated hazard function, ie,

$$\hat{\Lambda}_0(t) = \sum_{i:t_i < t} (d_i/n'_i)$$

and thus in the estimated survival function obtained by exponentiating the negative of this sum.

The structure of n'_i shows that this is exactly the principle implied by the Breslow estimator. Regardless of the value of z, or whether there are any subjects in the dataset with that specific covariate pattern value, the $\hat{S}_z(t)$ for z, calculated as a power of the step function $\hat{S}_0(t)$, has as many steps as there are distinct failure times in the overall dataset, regardless of the exactitude with which π_i is estimated.

If the values of some covariates in a dataset are located far from 0 (eg, ages, blood pressures, cholesterol levels), $\hat{S}_0(t)$ will be quite extreme relative to that for any covariate pattern in the dataset. The user does not usually look at this fitted curve, but rather at the $\hat{S}_0(t)^{\exp[\hat{\beta} z]}$ for each covariate pattern, z. However, if all covariates are centered, $\hat{S}_0(t)$ and the classic Kaplan-Meier curve will be close to each other, but, because of the nonlinearities involved, they will not be identical.

When he proposed his estimator, Breslow[3] also reported on “a covariance analysis of survival data arising from a clinical trial involving 268 patients on 5 regimens” that he had carried out. His estimate of $S_0[t]$ agreed with the more complicated estimate of Cox “to within 0.001 at each time point.” Given the large risk sets, the accuracy of his approximation should not have that surprising. But “even more surprising” to Breslow was the fact that “neither departed greatly from the unadjusted Kaplan-Meier estimate, obtained by setting $\hat{\beta} = 0$ in the expression for $\hat{\pi}_i$ above,” and that “this was true in spite of the fact that the covariate had a marked effect on survival.” In view of the above, possible reasons are that the variables were centered, or that, at the very least, the (scalar) values of the linear predictor derived from them were approximately centered.

From what I can discern, the phreg procedure in SAS (SAS Institute, Cary, NC) and stcox in Stata (Stata Corp, College Station, TX) use the Kalbfleisch and Prentice estimates of “α” (the complement of π) by default, while the basehaz function in the survival package in R (www.r-project.org) uses the Breslow estimate of π. If the risk sets are at all large, and the ties relatively few, then the differences between the results of the various approaches will be minor. Even with these minor differences in approach, all of them use synthetic (model-based) denominators made up of “baseline-equivalent” subjects created along the lines described above.

REFERENCES

1. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Ser B*. 1972;34:187–220.
2. Collett D. *Modelling Survival Data in Medical Research.* 2nd ed. Chapman & Hall/CRC; 2003.
3. Breslow NE. Contribution to the discussion of paper by D.R. Cox. *J R Stat Soc Ser B*. 1972;34:216–217.
4. Kalbfleisch JD, Prentice RL. Marginal likelihoods based on Cox’s regression and life table model. *Biometrika*. 1973;60:267–278.
5. Hanley JA. “Transmuting” women into men: Galton’s family data on human stature. *Am Stat*. 2004;58:237–243.